Kartu-Verbs: A Semantic Web Base of Inflected Georgian Verb Forms to Bypass Georgian Verb Lemmatization Issues

Mireille Ducassé

Kartu-Verbs: A Semantic Web Base of Inflected Georgian Verb Forms to Bypass Georgian Verb Lemmatization Issues

Ducassé Mireille[1]

[1] Univ Rennes, INSA Rennes, CNRS, IRISA - UMR 6074, France, mireille.ducasse@irisa.fr

Abstract

The Georgian language has a complex verbal system, both agglutinative and inflectional, with many irregularities. Inflected forms of a given verb can differ greatly from one another and it is still a controversial issue to determine which lemmas should represent a verb in dictionaries. Verb tables help people to track lemmas starting from inflected forms but these tables are tedious and error-prone to browse. We propose Kartu-Verbs, a Semantic Web base of inflected Georgian verb forms. For a given verb, all its inflected forms are present. Knowledge can easily be traversed in all directions: from Georgian to French and English; from an inflected form to a *masdar* (a verbal noun, the form that comes closest to an infinitive), and conversely from a *masdar* to any inflected form; from component(s) to forms and from a form to its components. Users can easily retrieve the lemmas that are relevant to access their preferred dictionaries. Kartu-Verbs can be seen as a front-end to any Georgian dictionary, thus bypassing the lemmatization issues.

Keywords: Georgian verbs; Inflected forms; Dictionary front-end; Semantic web tool; Prolog

1 Introduction

Georgian is a Caucasian language, mother tongue of about 5 million people. It has its own alphabet.[1] Georgian grammar has a complex verbal system. Some issues are illustrated below from a beginner's perspective (for more details see for example (Anderson 1984; Tuite 1998; Assatiani and Malherbe 2011; Gérardin 2016). There are numerous irregular verbs and the language is both agglutinative and inflectional. Conjugation can modify both the beginning and the ending of verbs. For example, verb "to work" (mushaoba - მუშაობა), at the first person plural of present tense gives "vmushaobt" (ვმუშაობთ). Note the "v" at the beginning of the verb to mark the first person, and the "t" at the end to mark the plural. Some tenses, such as future, often introduce a preverb. For example, for verb "to work", the first person singular future is "vimushaveb" (ვიმუშავებ). An "i" has been inserted after the "v" marker of first person; it is somewhat regular for a large set of verbs. Note that "ob" has changed into "eb"; it is typical of a smaller subset of verbs. The apparition of post-radical "v" is more exceptional. Many verbs have different stems at different tenses. For example, the third person singular forms of verb "to see" are respectively "khedavs" (stem, "khed", ხედ) at present and "nakha" (stem, "nakh", ნახ) at aorist. Indications of directions are given by prefixes. For example, "I go" = "mivdivar" (მივდივარ), "you go down" = "chadikhar" (ჩადიხარ), "she comes" = "modis"(მოდის). Note the different prefixes, and the very different markers of persons. A dozen prefixes can be used. Table 1 gives the conjugation tables for 3 tenses, present, future and present perfect. Beyond the above comments, note the mechanism at present perfect, the preradicals are very different. Those preradicals are used at other tenses, for other groups. For example the 3 persons singular at present of group 2 verb "to love" ("siqvaruli", სიყვარული) are respectively, "miqvars" (მიყვარს),"giqvars" (გიყვარს) and "uqvars" (უყვარს).

To work /მუშაობა		Present		Future		Present perfect	
I	მე	vmushaob	ვმუშაობ	vimushaveb	ვიმუშავებ	mimushavia	მიმუშავია
you	შენ	mushaob	მუშაობ	imushaveb	იმუშავებ	gimushavia	გიმუშავია
s.he	ის	mushaobs	მუშაობს	imushavebs	იმუშავებს	umushavia	უმუშავია
we	ჩვენ	vmushaobt	ვმუშაობთ	vimushavebt	ვიმუშავებთ	gvimushavia	გვიმუშავია
you	თქვენ	mushaobt	მუშაობთ	imushavebt	იმუშავებთ	gimushaviat	გიმუშავიათ
they	ისინი	mushaoben	მუშაობენ	imushaveben	იმუშავებენ	umushaviat	უმუშავიათ

Table 1: Three tenses of verb "to work/მუშაობა".

The preceding examples are not exhaustive. They only aim at illustrating the difficulty of morpho-syntactic analysis of Georgian verbs and pave the way to introduce some issues of verb lemmatization in Georgian Dictionaries (details can be found in (Margalitadze 2020; Gippert 2016). Georgian has no infinitive. Most dictionaries use the "*masdar*" (a verbal noun that is the form closest to an infinitive) as lemma to represent a verb.[2] However, for neophytes, going from a conjugated form to a *masdar* can be a real challenge. For example, for "chamodikhar" (ჩამოდიხარ, "you come down"), the *masdar* is "mosvla" (მოსვლა, "coming"). Many projects give samples of inflected forms as lemma(s). For example, third person singular future is used in (Daraselia and Sharoff 2016). The "Comprehensive Georgian-English Dictionary" presents, for all verbs, *masdar and 3rd person singular in present and future tenses, both active (transitive) form and*

[1] We use a transliteration in Latin characters, in this article and in Kartu-Verbs, to ease non-native Georgian speaker's reading. The transliteration is currently "French" oriented for historical reasons.

[2] In our system and in this report, a *masdar* is currently improperly called "Georgian infinitive" because it is easier to understand for the (French or English non-linguist) target users.

passive (intransitive) form, with markers for the indirect object in the third person. This is more exhaustive than in any previous bilingual Georgian dictionary (Rayfield et al., 2006). It is, however, still difficult for a neophyte to track the above-mentioned "chamodikhar". The Georgian-German dictionary of Tschenkéli et al. (1965) uses *the abstract verbal root under which all subparadigms are listed. It can result in an extremely complex structure of entries* (Gippert 2016). While this representation is very informative for linguists, it is too cumbersome for beginners, especially as many roots consist of only one or two characters.

Some linguists provide exhaustive tables of inflected forms, for example the Georgische Verbtabellen of (Chotiwari-Jünger et al. 2010) or the "Biliki series" books by Nana Shavtvaladze.[3] The latter contain conjugation tables of several types in appendix of the lessons. The first type of tables (called "whole conjugation tables" in the following) concerns the verbs introduced in a given lesson. They are systematically conjugated at all the tenses that have been introduced in the lessons so far. In these tables, *masdar* and English translation are also given. The second type of tables (called "sample tables" in the following) gives a list of conjugation samples, one line per verb. A line contains firstly an English translation and, for each tense introduced in the book, an inflected form at the third person singular. Those tables contain invaluable information; they are a tremendous help for neophytes. However, learners have to browse through different books to find relevant information. Finding an inflected form (in Georgian) in order to translate from Georgian to English is difficult. Indeed, the lines are sorted by English translation. When searching for an inflected form, learners have to check each one of the more than 10 000 entries. Furthermore, the inflected forms use the Georgian alphabet that is a big hurdle for beginners. Exceptions, which are quite common, cannot always be anticipated from the sample tables. Verbs introduced in the first books do not have a complete "whole conjugation" table because few tenses have been presented at the time these verbs are introduced. Searching is thus tedious, it takes time and it is not granted that users find an entry.

We propose Kartu-Verbs, a Semantic Web base of Georgian inflected verb forms that can be seen as a front-end to any dictionary, thus bypassing the lemmatization issues.[4] When a verb is in the base, all its inflected forms are present and users can retrieve the lemmas relevant to access their preferred dictionary. As illustrated in depth in Section 2, knowledge can easily be traversed in all directions: from Georgian to French and English and conversely; from an inflected form to a *masdar* and from a *masdar* to any inflected form; from component(s) to forms and from a form to its components. In order to build the base, conjugation rules, taking exceptions into account, are built in Prolog, a programming language designed for language processing (Colmerauer 2011). The generated forms are integrated within a Semantic Web tool, Sparklis, which can retrieve information from their facets, and which allows users to smoothly refine their queries by giving them suggestions (Ferré 2017). The base currently contains over 15 000 inflected forms related to 278 verbs for 10 tenses.[5] As discussed in Section 3, in comparison with related work and to our best knowledge, our tool is the only one of its type.

2 Using Kartu-Verbs, Our Georgian Verb Form Base

This section illustrates how to use our base of Georgian verb inflected forms and demonstrates the power of the tool. As the base is primarily meant to be a companion of the "Biliki" books already mentioned, we use the knowledge structures defined by Nana Shavtvaladze (groups, subgroups, morphological decomposition, etc). We are aware that the morphological decomposition is simplified, for example with respect to the work of Kevin Tuite (1998). Section 2.1 illustrates how to find information about an inflected form, our initial goal for the project. Section 2.2 shows that it is equally easy to get information starting from an English infinitive. Section 2.3 shows how to build a sample of conjugation. Section 2.4 describes how to gain conjugation information from a given ending. Sections 2.5 and 2.6 illustrate how to gain knowledge by comparing similar forms or stems. Section 2.7 shows how to check hypotheses about preradicals using logical operators. Section 2.8 discusses more sophisticated queries to gain meta-knowledge about the base using aggregates.

2.1 Finding Information About an Inflected Form

The user interface of Kartu-Verbs consists of 3 areas related respectively to the query, the suggestions and the results. Figure 1 shows two of those areas: the query area on the left-hand side and the "Suggestions" area on the right-hand side. The displayed query enables to find 12 features of inflected forms: its form in Georgian alphabet, person, number, tense, ..., French infinitive. The "Suggestions" area is itself divided in two areas. On the left, the "Types and Relations" area suggests features that can still be added to the query; on the right, the "Identities or Values" area suggests some of the verb forms that match the query. Let us assume that the user is interested in the "inadirebdnen" verb form and that he would also like to have information about its "Georgian infinitive" feature. He can click on both suggestions. Figure 2 displays the query and result areas after those selections. The query has been automatically updated. At the top there is no longer "give me every verb" but "inadirebdnen", and "Georgian infinitive" has been added in the list of features. In the result area the, now, 13 requested features about "inadirebdnen" correspond to 13 columns. We can see, for example, its form in Georgian alphabet, "ინადირებდნენ", and that it corresponds to the third person plural of both conditional and future conjunctive.

Any field could have been used to search the base. As opposed to paper tables, there are no predefined uses. As illustrated above, all queries are built using suggestions. Users do not have to invent anything. They can use filters to help Sparklis

[3] Biliki, Georgian Language For English Speakers. See http://lsgeorgia.com.

[4] The base is available at https://www-semlis.irisa.fr/software/georgian-verb-inflected-forms-base/

[5] Present, imperfect, conjunctive, future, conditional, future conjunctive, aorist, optative, present perfect, past perfect

propose relevant suggestions, then queries are built solely by clicking on suggestions that are necessarily relevant. The benefits are threefold, firstly it is easier to find something in a list than typing it, secondly users cannot mistype, and lastly, as a direct consequence, the queries can never give an empty result. That is a very strong property.

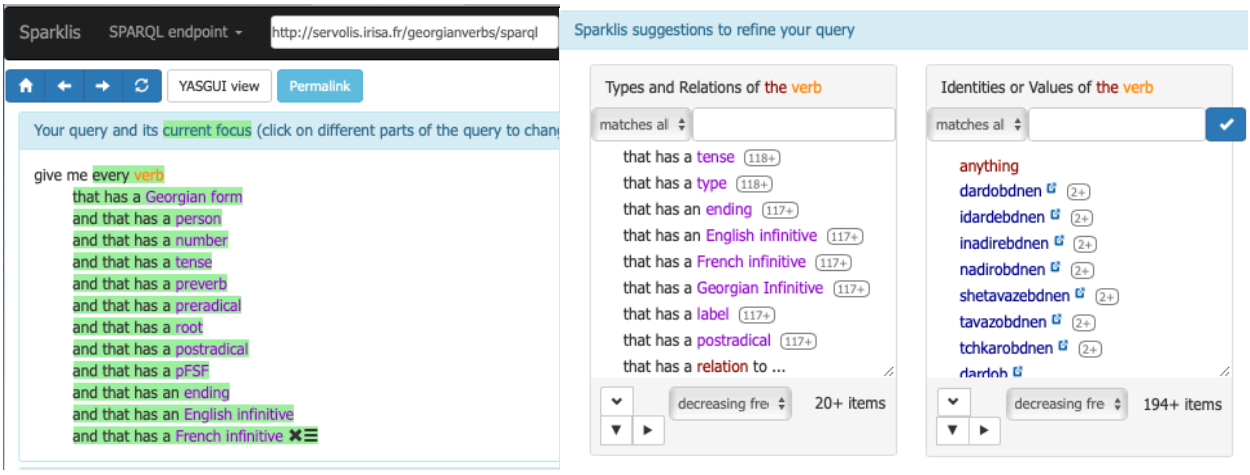


Figure 1: Defining a query with the help of suggestions.

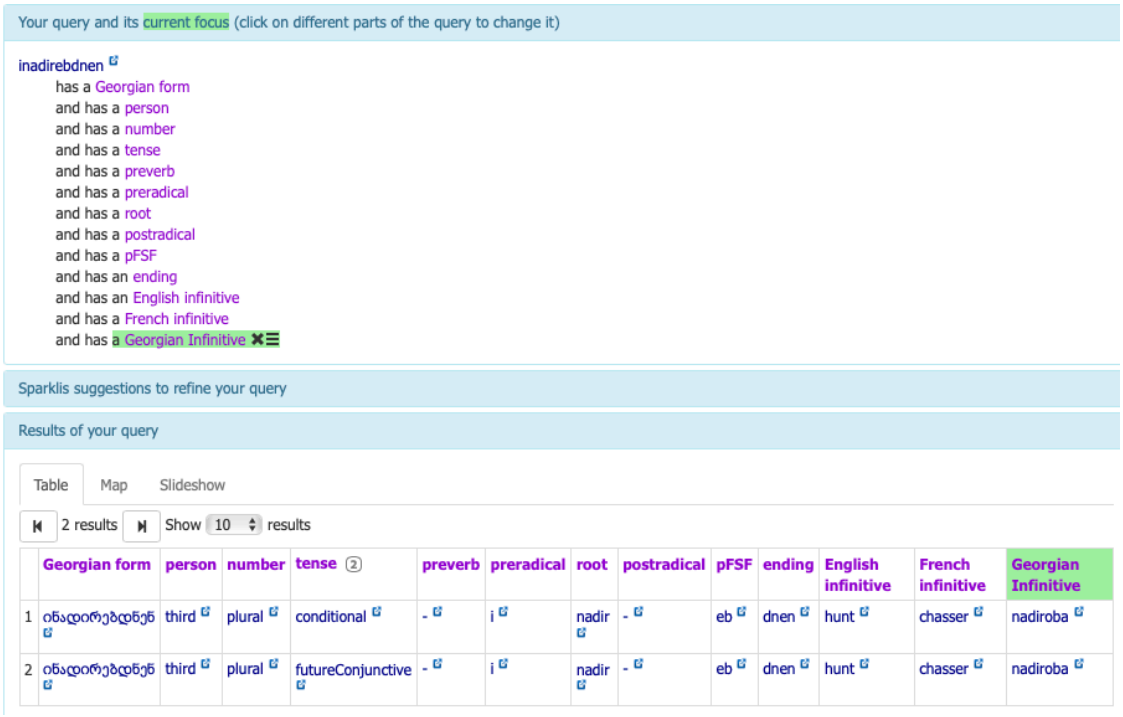


Figure 2: Getting information about an inflected form.

2.2 Finding Information From an English Infinitive

Figure 3 shows how to use criteria to search for information. Values are given to some features: the English infinitive has to be “to live”) and the tense has to be “present”. The features that do not have a value specified in the query (here “Georgian form”, “person”, “number”, “ending” and “Georgian infinitive”) are called “open”. They correspond to requested information and they produce the columns of the result area. The table in the result area gives the conjugation at the 6 persons at the present tense, with the requested information, in particular that it corresponds to Georgian verb “tskhovreba”. The features with a value in the query (here “English infinitive” and “Tense”) are not repeated in the result area. One needs the query in order to interpret the results. To get all tenses, users can specify them in turn in the query in order to get them one after the other, or they can leave the tense feature “open” and all tenses will be given. This view is equivalent to a “whole conjugation table” of the Biliki books mentioned in the Introduction. Note that, at any moment, users can add or retract any value or feature from the query. The result and suggestion areas are then automatically updated by Sparklis.

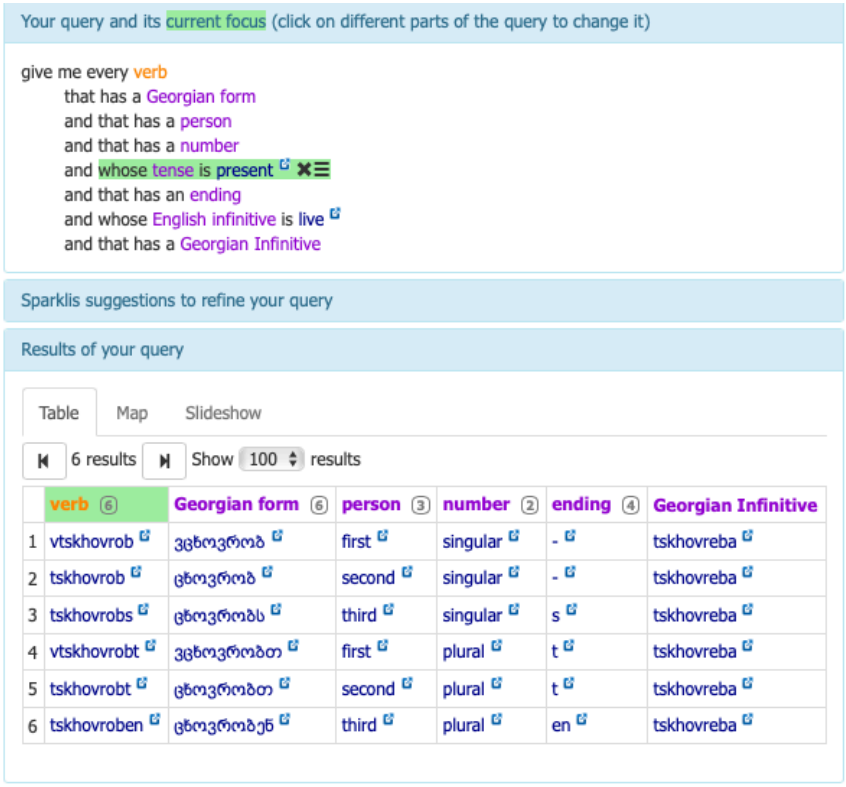


Figure 3: The conjugation of the six persons at a given tense for a verb given in English.

2.3 Building a Sample of Conjugation

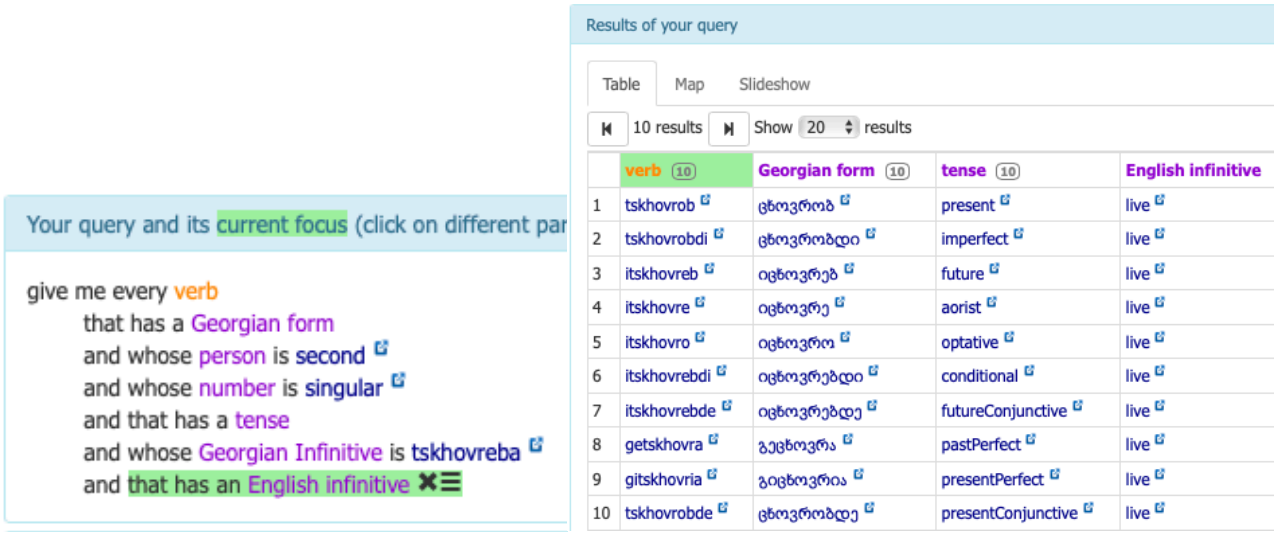


Figure 4: All 10 tenses at the second person singular of a verb given in Georgian.

Figure 4 shows how to conjugate Georgian verb “tskhovreba” at all the tenses known by the base for the second person singular. This is equivalent to a “Sample table” of the “Biliki” books. The advantage is that users can chose the person(s) they want or any criteria. Note that this time, we have specified the verb by its Georgian infinitive but we could have given one of its English or French infinitives.

2.4 Finding Possible Tenses From a Given Ending

Let us, now, assume that the user searches a verb form that is not present in the base but for which the user thinks that the ending is “da”. It could already be interesting to know the possible tenses. Figure 5 shows a query that sets the ending and asks for many features (“Georgian Form”, person, number, tense, preverb, PFSF and English infinitive), in order to try to map the searched verb to what is currently in the base. The result area shows 10 forms out of the more than 200 ones that match the query. The forms on display all correspond to a third person singular, at imperfect or conditional and with a PFSF being either “eb” or “ob”. It gives interesting trends. The user can check the remaining forms in order to confirm them (not illustrated here).

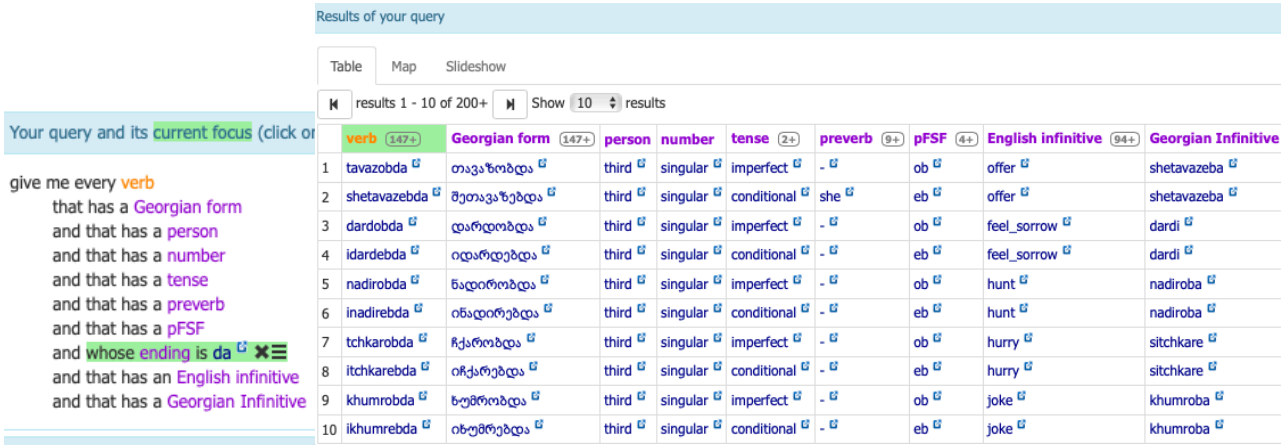


Figure 5: Finding possible tenses from a given ending.

2.5 Comparing Similar Forms

When learning, it is often useful to confront similar forms. For example, let us assume that the user realizes that he is confused about “to have someone” and “to resemble”. Figure 6 illustrates how to display the third singular present form for both verbs, using feature “French infinitive” and the “or” logical operator. The result area shows that the difference between the two forms consists in only one character. Thus, the user has learnt that “to have someone” has a “q” (“ყ”) as second letter and “to resemble” a “g” (“გ”).

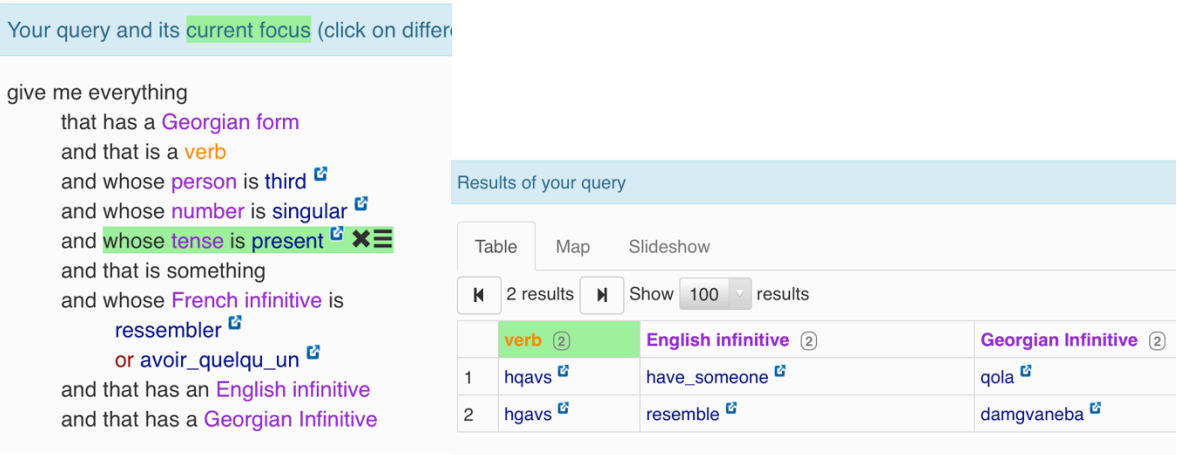
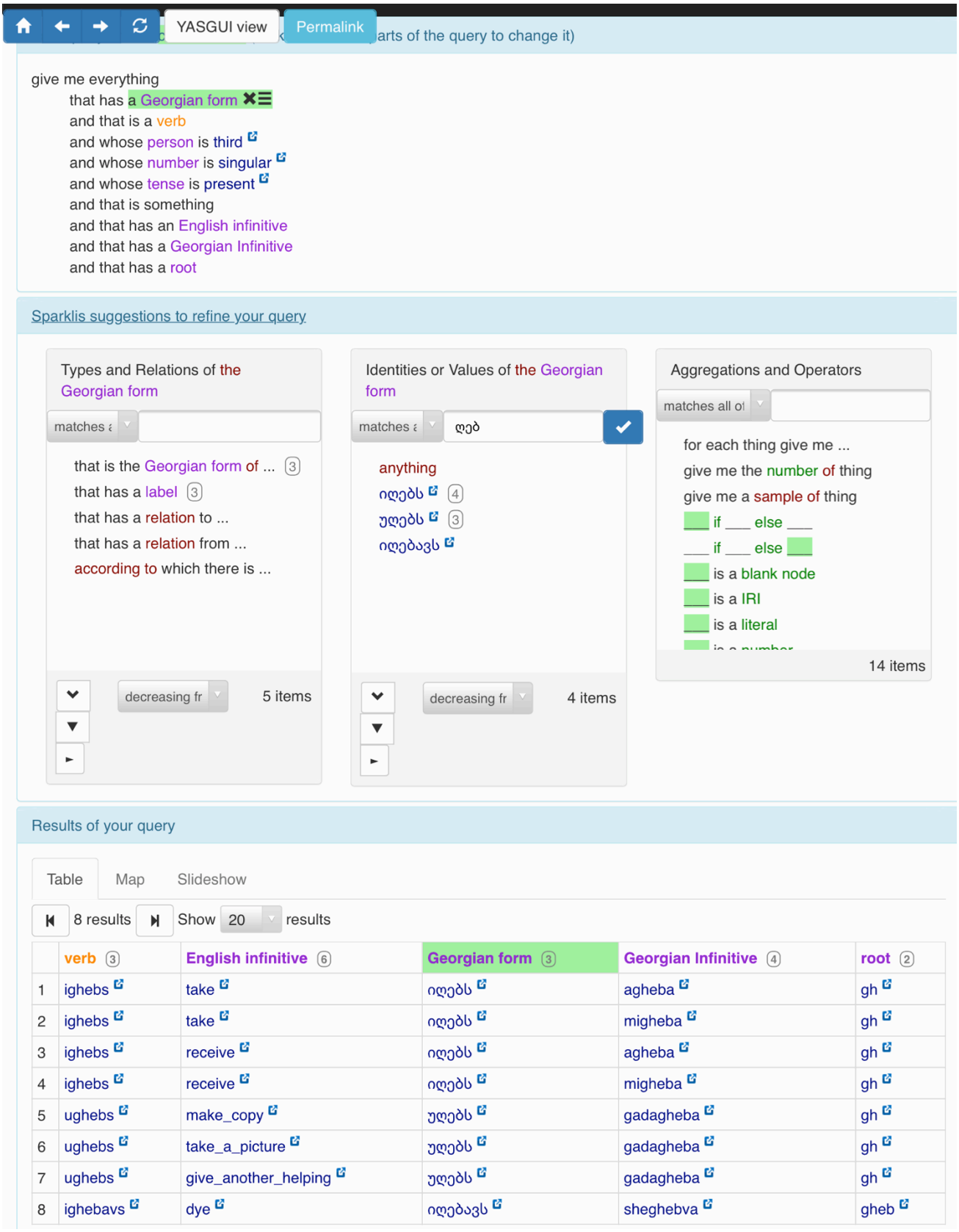


Figure 6: Comparing two similar forms.

2.6 Investigating Similar Stems

Similarly, let us assume that the user is confused about verbs containing “gheb” (“ღებ”) in their form, not knowing exactly which type of morpheme it is. Figure 7 illustrates how to use the suggestion area to help on this matter. The query requests the verb to be third singular present and asks information about English and Georgian infinitives as well as stem/root. The green underlining in the query area indicates that the focus for the suggestions is on the “Georgian form”. The user has typed “ღებ” in the suggestion area and Sparklis has automatically produced 3 suggestions (“იღებს”, “უღებს” and “იღებავს”). The result area shows 8 results for verbs whose Georgian form at third singular present matches “ღებ”. For verb “to dye” the stem/ root is exactly “gheb” and the PFSF is “av”, while for verbs of the “to take/receive” family the stem/root is “gh” and “eb” is the PFSF. Thus, the user has learnt that “to dye” and “to take/receive” are not acquainted. Note that “იღებს” and “უღებს” each give several answers. “იღებს” corresponds to 2 different Georgian infinitives and 2 different English translations.[6] “უღებს” corresponds to 3 different English translations. Figure 7 shows the complete display of the “Suggestions” area. The “Types and Relations” area (on the left) and the "Identities or Values" area (in the middle) have already been introduced. Let us remind here that they suggest, respectively, features that can still be added to the query and some of the values that match the query. The “Aggregation and Operators” area (on the right) allows users to build more sophisticated queries as illustrated in the following sections.

[6] Actually, “agheba” means “to take” and “migheba” means “to receive”. There should be 2 lines for “ighebs” and not 4. While it is a powerful feature to be able to display several lines for a given inflected form, the system shall be enhanced to remove irrelevant products.



	verb 3	English infinitive 6	Georgian form 3	Georgian Infinitive 4	root 2
1	ighebs	take	იღებს	agheba	gh
2	ighebs	take	იღებს	migheba	gh
3	ighebs	receive	იღებს	agheba	gh
4	ighebs	receive	იღებს	migheba	gh
5	ughebs	make_copy	უღებს	gadagheba	gh
6	ughebs	take_a_picture	უღებს	gadagheba	gh
7	ughebs	give_another_helping	უღებს	gadagheba	gh
8	ighebavs	dye	იღებავს	sheghebva	gheb

Figure 7: Investigating similar stems.

2.7 Learning About Preradicals

Let us assume that the user believes that a first person always has preradical “v” or “vi”. Figure 8 shows a query, using logical operators “and” and “not”, that searches for forms at first person (singular or plural, as number is not specified) and whose preradical is neither “v” nor “vi”. The addition of “something” in the query tells the system that preradical values are of interest. The suggestion area immediately shows that there are at least 6 other possibilities. For example, as illustrated in Table 1 “mi” and “gvi” are used at present perfect for some verbs. The user has to refine his knowledge!

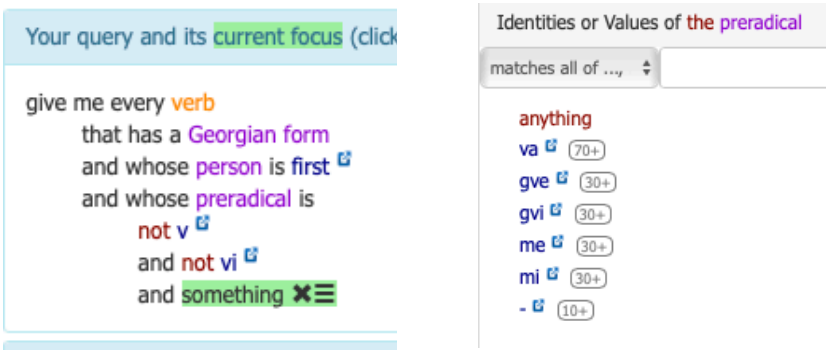


Figure 8: Learning preradicals at first person, using logical operators “and” and “not”.

2.8 Meta-Knowledge About the Base

	group	sample of verb	number of verb	number of pFSF	number of preverb	number of Georgian Infinitive
1	g4	khar	1626	13	18	32
2	g2	mqopnis	599	3	7	11
3	g3	movip'ove	3051	7	8	56
4	g1	vtavazob	9988	15	13	179

Figure 9: Gaining knowledge about the group distribution, their numbers of PFSF, preverbs and Georgian infinitives using aggregation operators.

Sparklis also enables to gain knowledge about the current base. For example, Figure 9 shows a more sophisticated query, using aggregates, to gain information about the distribution of the 4 Biliki groups. The query tells that the features of interest are group, ending, pFSF, preverb and Georgian infinitive. It requests that all the verbs with these features are grouped according to their group (“g1” to “g4”). For each group, a sample should be given and the number of verb inflected forms, the number pFsF, the number of preverbs and the number of Georgian infinitives are requested. On the result area, we can see that group g1 is the largest one. In the current state of the base, it gathers 179 Georgian infinitives, 9 988 inflected forms, 15 different PFSF and 13 preverbs. The result area also shows a sample of each group. Note that this query, as all the previous ones, was built solely by clicking. Here the right hand part of the “Suggestions” area had been used (see previous section).

3 Discussion, Perspectives and Conclusion

To our best knowledge, our tool is the only one of its type. We have, for example, found nothing specific to the Georgian language on the MultiTAL platform,[7] expert in automatic language processing (TAL) focused on Eastern and/or poorly endowed languages (Sadoun et al. 2016). The Georgian Wiktionary[8] is aimed at Georgian speaking people. It is of no help to people who are beginning to learn the language. Google translate[9] is still doing quite poorly to translate Georgian verbs. INESS::XLE-Web,[10] the system of Paul Maurer (2007), is dedicated to linguists. It is able to parse sentences and produce syntax tree of a number of languages, including Georgian. While its linguistic power is much larger than what Kartu-Verbs offer, the information that we need is buried in the syntax tree and not really accessible to beginners. Furthermore, there are no transliterations, no translations, and last but not least, none of our querying possibilities.

Our project is still under development. Currently, the base contains over 15 000 inflected forms related to 278 verbs for 10 tenses (present, imperfect, conjunctive, future, conditional, future conjunctive, aorist, optative, present perfect, past perfect). The forms have been generated and tested by students who are native Georgian speakers. At least all the verbs of the "Biliki" books are covered. One can expect that the most useful verbs for every day's life are already present. According to Tuite (1998), 5 tenses are missing: present iterative, imperative, permansive, mixed conjunctive present, perfect conjunctive.

The short-term perspectives are as follows. The current verb forms are being systematically tested. We are still in the process of analyzing exceptions and irregularities. A library of usual queries is under construction. Phonetic and English-oriented transliterations are planned in order to help non-French users. More verbs will be added. Links to an actual electronic dictionary will be inserted (for example, to the Comprehensive Georgian-English Dictionary by D. Rayfield, on the site of the National Parliamentary Library of Georgia. [11])

In the medium term, we have to slightly revise the ontology that is structuring our form description in order to use a vocabulary more standard in linguistics and to be able to accommodate other types of words (nouns, adjectives, ...). We have to adapt the generation rules in order to be able to build forms with direct and indirect object markers (see for example (Assatiani and Malherbe 2011)), a feature that is especially confusing for French and English speakers. For example, "I do" (without other indication) ="vak'eteb"(ვაკეთებ), "I do for me" = "vik'eteb"(ვიკეთებ), "I do for you" ="gik'eteb" (გიკეთებ) [note the disappearance of the first person marker,"v"], "you do for me"="mik'eteb"(მიკეთებ), etc.

In a longer term, we want to complete the system to help users 1) enter new verbs, 2) validate the newly produced inflected forms and 3) update the conjugation program when exceptions are detected by experts. At some point, it will be important to ensure that the tool is collaborative, and that any user can suggest modifications and new entries in the database in a safe way.

Kartu-Verbs is an on-going project with many perspectives. In its current state it is already a successful proof of concept. In this paper we have shown how versatile and powerful its querying mechanisms are and how they can help users to easily get information about verbs that they encounter in Georgian text whatever their form. Kartu-Verbs can be used as a front-end to any Georgian dictionary, whatever lemmatization principles that dictionary uses for verbs.

4 References

Anderson, S. R. (1984). On representations in morphology case, agreement and inversion in Georgian. Natural *Language & Linguistic Theory*, 2(2):157–218.

Assatiani, I. and Malherbe, M. (2011). *Parlons Géorgien*. L'Harmattan.

Chotiwari-Jünger, S., Melik′išvili, D., and Wittek, L. (2010). *Georgische Verbtabellen*. Buske.

Colmerauer A. (2011). From natural language processing to Prolog. *Presentation at the Academy of Mathematics and Systems Science*, Beijing, April 8, 2011. http://alain.colmerauer.free.fr.

Daraselia, S. and Sharoff, S. (2016). Enriching Georgian dictionary entries with frequency information. In Margalitadze, T. and Meladze, G., editors, *Proceedings of the 17th EURALEX International Congress*, pages 321–327, Tbilisi, Georgia. Ivane Javakhishvili Tbilisi University Press.

Ferré, S. (2017). Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language. *Semantic Web: Interoperability, Usability, Applicability*, 8(3):405–418.

Gippert, J. (2016). Complex morphology and its impact on lexicology: the Kartvelian case. In Margalitadze, T. and Meladze, G., editors, *Proceedings of the 17th EURALEX International Congress*, pages 16–36, Tbilisi, Georgia. Ivane Javakhishvili Tbilisi University Press.

Gérardin, H. (2016). *Les verbes intransitifs primaires et dérivés en Géorgien : Description morphosyntaxique, sémantique et dérivationnelle*. PhD thesis, Institut National des Langues et Civilisations Orientales. UMR 7192 – « Proche-Orient-Caucase : langues, archéologie, cultures ».

Margalitadze, T. (2020). Lexicography of Georgian: a brief overview. *Language@Leeds Working Papers in Linguistics*, 1. Forthcoming.

Meurer, P. (2007). A computational grammar for Georgian. In *International Tbilisi Symposium on Logic, Language, and Computation*, pages 1– 15. Springer.

[7] http://multital.inalco.fr

[8] https://meta.wikimedia.org/wiki/List_of_Wiktionaries_by_language_family - Kartvelian_lan- guages_(8,074_-_1_-_1)

[9] https://translate.google.fr

[10] http://clarino.uib.no/iness/xle-web

[11] დიდი ქართულ-ინგლისური ლექსიკონი. http://www.nplg.gov.ge/gwdict/index.php?a=index&d=46

Rayfield, D., Apridonze, S., Chanturia, A., Amirejibi, R., Broers, L., Chkhaidze, L., and Margalitadze, T., editors (2006). *A Comprehensive Georgian-English Dictionary*. Garnett Press.

Sadoun, D., Mkhitaryan, S., Nouvel, D., and Valette, M. (2016). The MultiTal NLP tool infrastructure. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities* (LT4DH), pages 156–163.

Tschenkéli, K., Marchev, Y., and Flury, L. (1965). *Georgisch-Deutsches Wörterbuch*, volume 2. Amirani-Verlag Zürich.

Tuite, K. (1998). *Kartvelian morphosyntax: Number agreement and morphosyntactic orientation in the South Caucasian languages*. Lincom Europa Munich.

Acknowledgements

We are thankful to Irma Grdzelidze and Tina Margalitadze from Ivane Javakhishvili Tbilisi State University for providing more hindsight of this work. We are indebted to Tina for pointing out the verb lemmatization issue. Thanks to the Erasmus+ International Credit Mobility program, the following Georgian students contributed to the project: Keti Meipariani, Mariam Asatiani, Mikheil Maisuradze, Tamari Kldiashvili, Tamar Sharabidze, Beka Chachua, Ana Idadze, Veriko Nikuradze, Tornike Tchanturia, Ana Elchishvili, and Aleksandre Jajanidze. We are grateful to Keti Meipariani who managed the working teams of two semesters and who remotely helps the following teams. Last but not least, we want to warmly thank Sébastien Ferré for his support to use Sparklis.